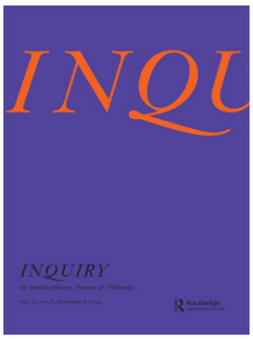
Artificial agents

Veluwenkamp, Herman; Hindriks, Frank

Published in:
Inquiry-An interdisciplinary journal of philosophy

DOI:
10.1080/0020174X.2024.2410995

Publication date:
2024



Routledge
Taylor & Francis Group

Inquiry
An Interdisciplinary Journal of Philosophy

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/sinq20

Artificial agents: responsibility & control gaps

Herman Veluwenkamp & Frank Hindriks

To cite this article: Herman Veluwenkamp & Frank Hindriks (03 Oct 2024): Artificial agents: responsibility & control gaps, Inquiry, DOI: 10.1080/0020174X.2024.2410995

To link to this article: https://doi.org/10.1080/0020174X.2024.2410995

Published online: 03 Oct 2024.

INQUIRY
https://doi.org/10.1080/0020174X.2024.2410995

Routledge
Taylor & Francis Group

Artificial agents: responsibility & control gaps

Herman Veluwenkamp and Frank Hindriks

Department of Ethics, Social and Political Philosophy, Faculty of Philosophy, University of Groningen, Groningen, The Netherlands

ABSTRACT

Artificial agents create significant moral opportunities and challenges. Over the last two decades, discourse has largely focused on the concept of a 'responsibility gap.' We argue that this concept is incoherent, misguided, and diverts attention from the core issue of 'control gaps.' Control gaps arise when there is a discrepancy between the causal control an agent exercises and the moral control it should possess or emulate. Such gaps present moral risks, often leading to harm or ethical violations. We propose a second-order 'duty of moral control' that mandates closing these gaps to reduce risks within acceptable moral limits. Our analysis encompasses both autonomous machines and collective agents, acknowledging their similarities and key differences in constitution and moral status. We suggest four methods to close control gaps: ensuring artificial agents attain moral agency, providing meaningful human control, implementing safety engineering, and employing social control. These methods aim to responsibly integrate artificial agents into society. We conclude that a realistic approach, which addresses the practical problems posed by control gaps, is essential. This approach provides solutions to manage the risks posed by artificial agents while maintaining acceptable moral standards, ensuring we responsibly harness their potential and address the ethical challenges they present.

ARTICLE HISTORY Received 15 August 2024; Accepted 26 September 2024

KEYWORDS Responsibility gaps; control gaps; moral agency; collective agents; autonomous machines

1. Introduction

Some of our worst nightmares feature machines or organizations that are out of control. Think, for instance, of George Orwell's *Nineteen eighty-four* (1949/2021), in which 'big brother is watching you' or of *The Terminator* (1984) in which a cyborg assassin begins his own killer spree. It is important to keep organizations and machines in check, in particular

CONTACT Herman Veluwenkamp h.m.veluwenkamp@rug.nl

insofar as they act autonomously. In the case of organizations, this means that the decisions they make and the actions they perform can, in a sense to be made precise below, come apart from those of their members. Machines do things on their own if they act without human supervision. The ways in which autonomous machines and organizations, or collective agents, differ from human beings raise complex moral questions, for instance concerning the attributing of moral responsibility. There is a growing awareness that progress is to be made by comparing them, which is what we do here (Bhargava and Velasquez 2019; Duijf 2022; Hakli and Mäkelä 2019; Laukyte 2017; List 2021; Singer 2013).

Thus far, responsibility and blame have been at the center of attention (Braham and van Hees 2018; Collins 2019; Duijf 2022; Himmelreich 2019; Köhler, Roughley, and Sauer 2017; Königs 2022; Matthias 2004; Munch, Mainz, and Bjerring 2023; Nyholm 2023; Pettit 2007; Placani and Broadhead 2024). Both autonomous machines and collective agents have been argued to be susceptible to responsibility gaps. In the case of autonomous machines, this is because human beings have been removed from the decision-making loop, which means that they can no longer be blamed. Collective agents do have humans in the loop, but a corporate wrongdoing can be due to an accumulation of infelicities too insignificant to warrant blame. In such cases, there seems to be a shortfall in responsibility or blame. As it is difficult to see how this could be, responsibility gaps have been taken to reveal an inconsistency in our moral framework (Copp 2007; Matthias 2004). In response, fatalists within responsible AI claim that we should refrain from deploying autonomous machines (Sparrow 2007). At the other extreme, quietists within social ontology maintain that responsibility gaps dissolve once organizations are regarded as suitable targets of blame (Copp 2007; Pettit 2007).

We argue, however, that there are no responsibility gaps.[1] Sometimes there is no blame to be attributed: what happened was merely an accident. In other cases, philosophers have been too quick to conclude that it is not possible to attribute blame. What has not been appreciated sufficiently is that human beings can be responsible indirectly. We go on to propose that the problem has been misidentified. In the kind of situations at issue, the agent exemplifies a shortfall in control. Furthermore, such 'control

[1]In this paper, we define responsibility gaps as shortfalls in responsibility. This interpretation is the predominant way of understanding these gaps. In other work, we have referred to this as the metaphysical conception of responsibility gaps (Veluwenkamp 2024). While there are other useful conceptions of responsibility gaps, these alternative views do not adhere to the notion that responsibility is deficient, which is the focus of this paper.

gaps' are dangerous, as they can lead to accidents or mistakes. They thereby expose others to risks that are morally unacceptable. Because of this, control gaps ought to be closed. Thus, control gaps present us with a practical challenge that neither fatalists nor quietists recognize.

Instead of focussing on responsibility and blame, we propose to concentrate on moral control and risk. In section 2, we introduce the notions of an autonomous machine and a collective agent. In sections 3 and 4, we discuss the notion of a responsibility gap and argue that they do not exist. And in section 5, we present the notion of a control gap and its relation to risk. Furthermore, we argue that control gaps ought to be closed. In section 6, we propose that this can be done, among others, by providing meaningful human control (Cavalcante Siebert et al. 2022; Horowitz and Scharre 2015; Santoni de Sio and Van den Hoven 2018). And we argue that, insofar as collective agents are concerned, this serves to secure their moral agency.

2. Artificial and moral agency

Autonomous Machines (AMs) and Collective Agents (CAs) are created and maintained by human beings. As such, they are artifacts. AMs are technological artifacts. Well known examples are self-driving cars and killer robots. In contrast, CAs are social constructs or social artifacts. They encompass corporations, universities and states. As we go on to discuss, AMs and CAs can plausibly be regarded as agents. Instead of biological agents, they are non-biological or artificial agents.[2] In order to see why and how they are agents, it is important to discuss the roles that human beings play in this context.

Ordinary machines are operated by human beings. What are known as 'semi-autonomous machines' are able to decide and act independently from human beings in some situations, but not all. Think, for instance, of sophisticated drones that are able to navigate through certain environments on their own. Yet, they cannot set their own goals. In this respect, they depend on human operators. AMs decide and act without human oversight and intervention. In this sense, they are autonomous. However, they are created and maintained by human beings. We call these human beings 'enablers', which we distinguish from 'operators.' They include designers and software engineers, manufacturers and regulators. AMs have enablers, but no operators.

[2]By 'artificial' we just mean 'human made.' Because of this, it can be used in relation to AMs as well as CAs. For an example as to how the term is used in relation to CAs, see Pettit (2017, 23–24).

In contrast, CAs have enablers as well as operators. Their operators are members who contribute *directly* to their decisions and actions. Think, for instance, of teachers, construction workers and police officers. Enablers can but need not be members. Compare, for example, a scheduler to a regulator. They facilitate the operators. Because of this, they are implicated in the actions of CAs only *indirectly*. As CAs have operators, it is more difficult to see why they might qualify as agents in their own right, as compared to AMs. One might think that their agency reduces to that of their members. However, it has been argued that CAs do act autonomously. In particular, due to the collective decision procedures they rely on, they can take decisions and perform actions that none of their members support directly.[3] Because of this, their agency is irreducibly collective (Pettit 2007; List and Pettit 2011).

But what kind of agency do artificial agents possess? We distinguish between minimal and moral agents. Minimal agents have mental states such as beliefs, desires or intentions, which are directed at, or about, something. Furthermore, they can reason, decide and act on their own (Himmelreich 2019). In addition to this, moral agents possess normative competence (Wallace 1994). This means that, apart from prudential matters, they can also respond to moral considerations. They can bring them to bear on their decisions and on their actions. John Fischer and Martin Ravizza (1998) make this point in terms of reason-responsiveness, which they break down into two abilities: to be receptive to moral reasons in thought, and to be reactive to them in action. We say that an agent who possesses these abilities has 'moral control.'[4]

AMs are typically seen as minimal agents. In contrast, collective agents are often regarded as moral agents.[5] This implies that CAs possess moral control, while AMs do not. Most of the time, we assume that this is indeed the case. However, we also explore the possibility that AMs are moral agents, while collective agents are minimal agents. In such cases, the comparison between AMs and CAs sheds light on possibilities that have remained under explored so far. In fact, our comparative approach serves as a heuristic tool, as it enables us to transpose assumptions from one domain to another. Thus, a distinctive feature of our

[3]The point can be made more precise as follows. The beliefs, desires and intentions of a collective agent do not only supervene on the beliefs, desires and intentions of its members, but also on its collective decision procedure (List and Pettit 2006).

[4]Fischer and Ravizza (1998) call this 'guidance control.' Because of this, we use the term 'moral control' instead as we prefer to remain neutral as to how exactly the kind of control that is necessary for moral responsibility is to be understood.

[5]For an example of authors arguing that AMs cannot be moral agents, see (Purves, Jenkins, and Strawser 2015) and (Véliz 2021). For a contrasting view, see (List 2021, 1229).

methodology is that we identify similarities and differences, so as to exploit inter-agential analogies and disanalogies. In this way, we promote a mutually beneficial interaction between research on artificial intelligence and social ontology.

Although human beings can form a useful point of reference, we do not assume that they set the standard, as is often done (Himmelreich 2019, 734; Sparrow 2007, 66). For instance, in some contexts, it is more insightful to compare artificial agents to young children, who are not yet full-blown moral agents. As we discuss below, it is also useful to compare them to semi-autonomous machines, which are not even minimal agents. For one thing, given the current state of technology, fully autonomous machines operating effectively across a wide range of areas are rare if not non-existent. Furthermore, as semi-autonomous machines have operators, they provide a bridge between AMs and CAs. We draw these analogies purposely, as our comparative approach stands to benefit from further differentiations. Yet, the core of the analysis we go on to present revolves around the distinction between minimal and moral agents. For one thing, we consider the possibility that responsibility gaps can be closed only if the artificial agent is a moral agent. Subsequently, we propose the notion of a control gap, which we define such that it can be applied to minimal as well as moral agents. In conclusion, we argue that the comparison is particularly fruitful when it comes to closing control gaps.

3. Responsibility gaps introduced

3.1. Moral responsibility

In order to bear moral responsibilities, an agent must be a moral agent and have moral control over their actions and the consequences of these actions. Moral responsibility can be forward-looking and backward-looking (van de Poel 2011). An agent bears a forward-looking responsibility if they have an obligation to do something. After the moment has passed at which the obligation should have been fulfilled, the agent is responsible for the action or outcome in the backward-looking sense. Depending on whether the obligation was met, the agent becomes a candidate for praise or blame.

Responsibility gaps concern blameworthiness.[6] An agent can be to blame for an action or its consequences. In either case, we will say the

[6]See Copp (2007) for a discussion of responsibility gaps in relation to forward-looking responsibility.

agent did something wrong. To be blameworthy for failing to fulfill an obligation, three conditions must be fulfilled. First, the agent must be a moral agent; second, they must have had causal control over the action or consequence, meaning that they must have performed the action or caused the relevant consequence; and third, they must have borne the right epistemic relation to it, such as knowing that what they did was wrong or that it had morally unacceptable consequences–or they *should* have known it (e.g. Hart 2008). Even though there is a lot of disagreement about how exactly these three conditions should be understood, there is widespread agreement that they have to be satisfied.

To determine whether the conditions are actually satisfied, it is important to check for defeaters: justifications, excuses and exemptions (Wallace 1994). Suppose that an agent has a *pro tanto* obligation to perform some action. An agent has a justification not to fulfill that obligation if it is overridden by another obligation. In that case, they do nothing wrong by not fulfilling the initial obligation, as long as they satisfy the overriding one. Someone can, for instance, justifiably violate their promise to help a friend move if their mother needs urgent care. Furthermore, an excuse is a situation-specific factor that defeats blame. Think, for instance, of dizziness, coercion and non-culpable ignorance. As these examples reveal, excuses can pertain to any of the three conditions of blameworthiness, which means that they can be agential, causal or epistemic. Finally, an agent is exempted from blame if it does not possess enough moral control. This can be due to some deficiency, such as psychopathy. But it can also be that a child has not yet acquired enough moral control to qualify as a moral agent.

Defeaters play an important role in section 3.3, where we present our diagnosis as to why people are sometimes too quick to conclude that there are responsibility gaps. Suppose that an agent has done something that is, in principle, wrong. But their responsibility is defeated. In that case, they are, in all likelihood, not to blame for what they did. We call this 'the no-blame point.' However, it could also be that they are to blame for it indirectly. For this to be the case, they must lack direct control over what they did, for instance because they were drunk at the time. Yet, they possess indirect control. Because of this, they could have prevented the mishap at an earlier point in time, for instance when they were still sober (McKenna 2008; Vargas 2005). This is 'the indirect-blame point.' Before we use these two points to criticize the notion of a responsibility gap, we first discuss what they are. This is important because there are no well-established definitions of responsibility gaps.

3.2. Responsibility gaps defined

A responsibility gap is a shortfall in responsibility. Suppose an artificial agent does something wrong or has bad consequences. A responsibility gap arises if it is impossible to blame anyone for the action or its consequences. But what makes it so difficult to attribute blame? Artificial agents are inherently complex. Autonomous Machines (AMs) often rely on intricate and opaque decision-making algorithms. Likewise, Collective Agents (CAs) can be complicated due to the way decision-making processes are structured and labor is divided.

But these are epistemic obstacles for attributing blame. And the notion of a responsibility gap has been introduced to capture a deeper problem, which concerns the roles that human beings play in relation to artificial agents (Copp 2007; Matthias 2004). Insofar as AMs are concerned, they do not have any operators who are causally implicated in their wrongdoings. Although CAs do have operators, their responsibilities might be defeated. In such situations, it seems that no human being can be blamed. Yet, this leaves open the possibility that blame is appropriate. If it is, there is a responsibility gap.

Thus, we define the notion as follows (Hindriks and Veluwenkamp 2023; Hindriks 2024a):[7]

> [RG] A responsibility gap exists exactly if (1) an artificial agent did something for which blame is fitting, even though (2) no human being can be blamed for it.

A responsibility gap can be 'technical' or 'collective,' depending on whether it is exhibited by an AM or a CA. We go on to argue that there is a striking difference between them.

3.3. Minimal and moral agents

A responsibility gap is problematic because no blame can be attributed even though it would be fitting to do so. As the notion of a gap suggests, solving the problem is a matter of closing the gap. This in turn is a matter of finding a way to attribute the blame after all. Now, suppose that the artificial agent that exhibits a responsibility gap qualifies as a moral agent. In that case, the blame can be attributed to it. Furthermore, doing so closes the gap. If, on the other hand, the artificial agent is a minimal agent, the blame really is unattributable. So, depending on

[7]The notion can also be defined so as to allow for degrees. In that case, the idea is that the blame that can be attributed does not meet the severity of the event.

whether the artificial agent is a minimal or a moral agent, the responsibility gap can be solved or is insoluble.

In light of the discussion in section 2, our working hypothesis is that AMs are minimal and CAs moral agents. This implies that collective responsibility gaps can be closed, while technical responsibility gaps cannot be. And technical responsibility gaps are indeed perceived as inescapable. If there is a responsibility gap, the unattributed blame is, so to say, lost. It 'evaporates' (Santoni de Sio and Van den Hoven 2018, 2). Because of this, technical responsibility gaps create an impasse: although blame is fitting, it is impossible to attribute it. Things would be rather different if AMs were moral agents. In that case, the blame could be attributed to them.

The situation is basically the reverse within social ontology. If CAs were minimal agents, collective gaps could not be solved. Yet, it is commonly assumed, in this context, that they are moral agents. Strikingly, David Copp (2007) starts from the assumption that CAs are minimal agents, which means that they cannot be blamed. Next, he hypothesizes that, in a particular case, the responsibilities of the operators are all defeated. They all have excuses or justifications. Hence, the operators are not blameworthy. However, it is rather implausible that nobody can be blamed for the wrongdoing. So, an impasse has been reached. To get out of it, Copp proposes that the CA should be regarded as a moral agent after all. In this way, the gap can be filled. This 'irreducibility argument' is a *reductio ad absurdum* of the claim that CAs are minimal agents.[8] As such, it supports the idea that collective agents are best seen as moral agents.[9]

If responsibility gaps can be solved, there is no reason to make a big fuss about them. This explains why philosophers within social ontology have adopted a quietist attitude towards responsibility gaps. To be sure, Pettit (2017) argues that, if collective agents were not moral agents, there would be responsibility gaps and they would create loopholes that human beings could use to escape blame. However, this is only a hypothetical problem.[10] In contrast, if responsibility gaps are insoluble, they must constitute a big problem. This explains the fatalist response within responsible AI, that we might have to refrain from

[8]In a somewhat similar vein, List and Pettit claim that, if the collective agent itself is not blamed, we would be 'allowing some responsible actions […] to go undetected.' (2011, 166)

[9]But does the irreducibility argument not generalize to AMs? This would mean that AMs must be moral agents such that blame can be attributed to them. The only reason why it might not generalize is the prior plausibility of moral agency of AMs, which is taken to be lower than that of CAs.

[10]See Bhargava and Velasquez (2019, 835) for a criticism of Pettit's argument.

employing AMs (Sparrow 2007).[11] Similarly, it might be argued that technical responsibility gaps create avenues for unscrupulous exploitation, where human creators of AMs can in fact escape blame for negative outcomes.

In sum, responsibility gaps are regarded as bridgeable or inescapable depending on whether the relevant kind of artificial agent is seen as a moral or a minimal agent. Furthermore, if they are soluble, the problem is merely a theoretical one. In contrast, if it is insoluble, the problem has important practical consequences. In the extreme, we might have to refrain from deploying artificial agents altogether. Against this, we go on to argue that the problem is misunderstood when put in terms of responsibility gaps. Instead, it concerns deficiencies regarding moral control. Furthermore, such deficiencies constitute not only a theoretical problem, but also a practical one. Thus, we stay clear from quietism as well as fatalism.

4. Against responsibility gaps

4.1. The incoherence argument

As we have reconstructed it, the notion of a responsibility gap combines two claims: blame is fitting, yet no one can be blamed. Here we propose that these claims are inconsistent. To this end, we present 'the incoherence argument':

1. An artificial agent did something for which blame is fitting (*RG*, condition 1).
2. No human being can be blamed for it (*RG*, condition 2).
3. If blame is fitting, there is reason to attribute it.
4. If there is reason to ascribe blame, it can be attributed.
5. Hence, if blame is fitting, it can be attributed.
6. Hence, (1) and (2) cannot both be true at the same time.

The key premise of this argument is that, if there is reason to ascribe blame, then it must be possible to do so (4). This is an instance of the claim that, if there is a reason to do something, it must be possible to do it. And this follows from the principle 'reason implies can' (Streumer

[11] Strikingly, Matthias draws a rather different conclusion: 'Still, we cannot do without such systems, because the pattern processing and systems control tasks that we must accomplish in our highly dynamic and complex environments are so complicated that they cannot be addressed by simpler, statically programmed machines.' (2004, 183)

2007). The subsidiary premise is that, if blame is fitting, there is reason to attribute it (3). As we use it, the term 'fitting' is inherently connected to the notion of a reason in this way. Thus, the 'reason implies can' principle entails that there can be no instances where we have a reason to ascribe blame without it being possible to do so. Hence, the idea that there is a gap or a remainder is untenable. The very concept of a responsibility gap is incoherent.

At first sight, this is a rather strong conclusion, especially given the widespread acceptance of responsibility gaps. To make it easy to assess the validity of the argument, we have reconstructed it here in terms of premises and conclusions. Elsewhere, we presented it more informally (Hindriks and Veluwenkamp 2023; Hindriks 2024a). Upon reflection, however, the conclusion is not as surprising as it might seem initially. Our practices of attributing moral responsibility to artificial agents have been characterized as 'incoherent' exactly because they involve responsibility gaps (see Copp 2007; Matthias 2004). This conclusion is more extreme than ours. We propose that, instead of our responsibility practices, the notion of a responsibility gap is incoherent.

Furthermore, the incoherence of the notion also explains why responsibility gaps are taken to have such strong and striking implications. If there are responsibility gaps, they should be closed. This is why responsibility gaps have been taken to imply that CAs should be considered moral agents (quietism). And, conversely, that if these gaps cannot be closed, that these artificial agents cannot be deployed (fatalism). However, not everyone who discusses responsibility gaps endorses these drastic conclusions. To convince them, we consider the debates about responsibility gaps in more detail.

According to what we call 'the blame argument', the no-blame point and the indirect-blame point, introduced in section 4.1, have not been appreciated sufficiently in discussions of responsibility gaps. Furthermore, the 'anthropomorphic mistake', mentioned in section 2, has it that it can be problematic to use human beings as a standard for how much blame is appropriate, as is often done. These diagnostic observations reveal why people are often too quick to conclude that there are responsibility gaps, even though there are none.

4.2. The blame argument

According to the blame argument, people can erroneously conclude that there is a responsibility gap if they fail to appreciate the no-blame point

and the indirect-blame point. The no-blame point is the claim that an action that is bad or wrong need not be blameworthy. People who do not appreciate this sufficiently may be too quick to conclude that blame is fitting. The indirect-blame point is the insight that, even if no one is to blame directly, someone might be to blame directly. If this insight is not appreciated enough, people may fail to see that blame can in fact be attributed. Suppose there appears to be a responsibility gap. The no-blame point puts pressure on the idea that blame is fitting (1). And the indirect-blame point challenges the idea that blame cannot be attributed (2). We go on to discuss these points in more detail so as to explain how they can explain why people sometimes mistakenly believe there is a responsibility gap.

The no-blame point concerns defeaters and their significance. Before affirming that blame is fitting, it is important to check if there are no defeaters. The reason for this is that it is tempting to move directly from the fact that something bad happened or that someone did something wrong to the conclusion that there is reason to blame someone. Yet, this does not follow if there is a defeater. Suppose someone did something that is morally bad. If they had a justification, which is an overriding obligation, what they did was not wrong. Consider next someone who did something that was wrong all things considered. If they have an excuse, they are not to blame for doing it. This also holds for an agent who is exempted and lacks moral control. As they are not a proper moral agent, they are not susceptible to blame. If someone claimed in such a situation that blame is fitting, they would have jumped to this conclusion. A more careful assessment would have revealed that there is no reason to attribute blame.[12]

Exemptions are relevant to Copp's irreducibility argument, discussed in section 3.3. The argument starts from the premise that the artificial agent is a minimal agent, which means that it lacks moral agency. This implies that the agent is exempted. So, the natural conclusion to draw is that blame is not fitting. Copp concludes instead that the premise that the agent is a minimal agent must be mistaken. And if it is a moral agent, then it can be blamed. In this way, the responsibility gap can be closed. Copp (2006, 220) maintains that the intuition that blame is to be attributed in the situation at issue is so weighty that the premise is to be rejected. But what if, for example, the collective agent is responsive

[12]List and Pettit fail to appreciate the no-blame point when they infer from the mere occurrence of some actions and consequences how much blame is appropriate. See Bhargava and Velasquez (2019, 834) for more on this.

only to prudential reasons and not to moral ones? In that case, no intuition about blame is weighty enough to justify the ascription of moral agency to the collective agent. So, Copp owes us an explanation of how his argument can accommodate exemptions.[13]

If an agent has an excuse, they do not bear direct responsibility. But an excuse can be traced back to an earlier event. In the case of the drunk driver, this is getting in the car while drunk (or failing to take measures to prevent this from happening). If the agent is to blame for this, they bear indirect responsibility and can be blamed after all. This is the indirect-blame point. It reveals that, in some cases, blame can be attributed even though there is a defeater. As we have argued elsewhere, the agent who is to blame need not be the same as the one who committed the wrongdoing. If you sabotage someone else's car, you may well be indirectly responsible for the ensuing accident (Hindriks and Veluwenkamp 2023). As we go on to discuss, this means that the indirect-blame point generalizes to artificial agents.

To explain how, we return to the distinction between operators and enablers. Operators contribute directly to the decisions and actions of artificial agents. Enablers facilitate artificial agents such that they can take decisions and perform actions. In particular insofar collective agents are concerned, they do so by enabling the operators. Now, suppose that an artificial agent does something bad. In principle, operators are directly responsible for this. In contrast, enablers bear indirect responsibility for it. But the operators might be excused, in which case their responsibilities are defeated. And perhaps the excuse can be traced to the enablers. If so, they are in principle to blame for what happened.[14]

Imagine an organization that operates a large industrial facility. Within this collective, there are various operators who oversee the daily operations and ensure that their sections comply with the overall safety standards set by the organization. However, due to an oversight, some crucial safety measures were not implemented by the enablers (e.g. the safety officers and facility managers). As a consequence, a mechanical failure occurs, leading to an injury to a worker stationed at the machine. Some of the operators will be implicated in the accident. However, they are excused. After all, the implementation and verification of such safety

[13]The alternative is to claim that there cannot be exemptions, as collective agents are necessarily moral agents. This claim also requires further defense (Hindriks 2024b).

[14]In practice, an employee can have both operative and enabling tasks. But in what follows we assume they are clearly separated.

measures fall under the responsibility of the enablers. Suppose that their responsibilities are not defeated. Then, the enablers are to blame for the harm, even if indirectly.

In a variant on this scenario, the artificial agent is a self-driving car. As it has no operators, the only agent who is directly implicated in the accident is the autonomous agent, which is a minimal agent. This implies that no one is directly responsible for the harm. Yet, it may well be that the enablers are to blame for it indirectly. Perhaps the accident is ultimately due to the negligence of the engineers.

The indirect-blame point is perhaps fairly obvious in the drunk driver case. However, it is, once again, more difficult to appreciate in relation to artificial agents. As just discussed, the agent to whom the wrongdoing is traced is not the same as the agent who committed it. What is more, the enabler(s) are usually nowhere near the place where it happens. Thus, it is not always immediately apparent who the responsible agents might be. Because of this, it is often worthwhile to look further before concluding that blame is not entirely attributable. There may be (other) enablers who are perfectly good candidates for blame. Checking for this is important so as not to prematurely conclude that blame cannot be attributed. In light of the above, we propose that, if someone claims there is a responsibility gap, it is likely they have failed to heed the no-blame point or the indirect blame point.

We end this section by discussing how the indirect-blame point just discussed reveals that a recent argument related to responsibility gaps fails. Pettit (2007, 113; 2017, 32) argues that the existence of a responsibility gap might create a loophole that individuals can abuse so as to escape blame. Similarly, the Expert Committee on human rights dimensions of automated data processing and different forms of artificial intelligence contends that the unchecked operation of AI-driven automation can lead to 'a serious responsibility gap, through which Big Tech reaps the benefits of these AI driven platforms without the concomitant burdens' (Yeung 2019). The worry is that Big Tech stands to profit from the technology's benefits while evading the moral responsibilities that come with it. In particular, social media companies have found an exploit allowing them the 'naked exercise of power without responsibility' (Yeung 2019)."

These arguments, however, are flawed. Consider a scenario where individuals intentionally set up things such that an artificial agent does something wrong, with the aim of escaping blame. In that case, they might not be directly responsible. However, exactly because they set out to use the alleged loophole, they are to blame indirectly. And this implies that there

is in fact no loophole (see also Bhargava and Velasquez 2019, 835). This is not to deny that AMs or CAs can abuse their power. Instead, the claim is that they cannot escape blame due to non-existent responsibility gaps.

In this way, a claim that is meant to illustrate how bad responsibility gaps are, helps to expose them as an unsupported if not confused idea. More generally, the no-blame point and the indirect-blame point explain why the notion of a responsibility gap has such a pull on us, which undermines the motivation for postulating it.

4.3. The anthropomorphic mistake

Philosophers compare artificial agents to human beings on a regular basis. In some cases, they do so to motivate the idea that blame is to be attributed in a particular situation or that something is amiss if this is not possible. For instance, Robert Sparrow argues that, if autonomous weapon systems inflict harm, these systems cannot be blamed for it. To explain why this is problematic, he observes that: '[h]ad a human being committed the act, they would immediately be charged with a war crime' (2007, 66). In a similar vein, List maintains: 'Crucially, however, there is no guarantee that the entirety of human responsibility will always be commensurate with the responsibility we would have attributed for an AI system's actions if those actions had been done by a human person.' (2021, 1227) The underlying idea is that the amount of blame which should be allocated should not depend on the kind of agent that is causally responsible for the harm. In both cases, human beings are seen as a model for artificial agents. In light of this, we call this 'the anthropomorphic argument', which will turn out to be mistaken.

The basic idea is that human beings form a relevant reference point for assessing the blameworthiness of artificial agents. It involves imagining that a human, not an artificial agent, committed a wrongdoing and assessing the appropriate level of blame in that hypothetical scenario. The key inference in the anthropomorphic argument is that the same amount of blame is fitting for the artificial agent. In this way, it provides further support for the claim that blame is fitting (1). It follows that, if blame cannot be attributed, there is a gap. In this way, the anthropomorphic argument is meant to support the existence of responsibility gaps.

Johannes Himmelreich goes as far as incorporating an appeal to anthropomorphic intuitions into his definition of responsibility gaps. They are situations where '(1) a merely minimal agent does *x*, such that (2) no one is responsible for *x*; but (3) had *x* been the action of a

human person, then this person would be responsible for *x*' (Himmelreich 2019, 734). But why would the comparison with human beings be relevant? Even if counterfactual blameworthiness settles actual blameworthiness, it is crucial to use the right counterfactual. And if artificial agents are minimal agents, it is more appropriate to compare them to young children, who also lack moral agency. This comparison suggests that it would be unproblematic if no blame were attributed, as young children are not blameworthy either. So, a proper analogy leads to the opposite conclusion.

Thus, the notion of a responsibility gap owes at least some of its popularity to an uncritical use of human beings as a point of reference. We refer to this as 'the anthropomorphic mistake.' Together with the blame argument, it forms our diagnosis as to why philosophers have concluded that there are responsibility gaps even though they do not exist. In fact, they cannot even exist, as the notion is incoherent. As the incoherence argument reveals, this is crucial if the notion is to support quietism or fatalism. All in all, the notion of a responsibility gap is surrounded by confusion. In light of this, we propose that we are better off without it. Yet, this does not mean that the situations at issue are always unproblematic. Instead, we go on to argue that, if enablers are to blame, the artificial agent exhibits a control gap.

5. Control gaps

The real problem concerns moral control. It might be that a moral agent is not sufficiently responsive to moral reasons. If so, it exhibits a control gap. Something similar holds for artificial minimal agents. Suppose that they interact with other agents or perform actions that have consequences for them. In that case, it would be good to construct them such that they act as if they are responsive to moral reasons. Or, as we will say, they should emulate a certain level of moral control. In light of this, we define the notion of a control gap as follows (Hindriks and Veluwenkamp 2023; Hindriks 2024a):

> **[CG]** An agent exhibits a control gap exactly if the causal control it actually has falls short of the moral control it should have or emulate.

To defend this definition, we argue that agents who exhibit a control gap pose an excessive risk to others (section 5.1). Because of this, they or their enablers have 'a duty of moral control' to ensure that there are no control gaps (section 5.2). In order to come full circle, we end by considering the

consequences that failing to fulfill this obligation has for backward-looking responsibility (section 5.3).

5.1. Risk

So, how much control should a moral agent have or emulate? One way of answering this question starts from the observation that control gaps are not restricted to artificial agents. Human beings can exhibit them as well. The famous drunk driver is a case in point. The alcohol he drank impaired his functioning. In particular, his ability to make decisions that are suitably receptive to moral reasons was compromised, as well as his ability to adequately react to them in a timely manner. This explains why, when he got into his car, the degree of moral control he had was lower than it should have been. However, this answer works for human beings. Perhaps it extends to other moral agents. But it is far from obvious that it is relevant to minimal agents. We stop short from concluding this, as it could involve us in another anthropomorphic mistake.

To make progress with this issue, consider another feature of the drunk driver case. Due to the control gap he exhibited, he exposed others to an excessive risk. It was so high as to be morally unacceptable. This is evident from the way we respond to cases like this, in particular when the risk materializes. To illustrate the difference between morally unacceptable and acceptable risk levels, we compare the drunk driver to 'the sober driver.' Suppose she also gets into an accident. And that this is not due to negligence or recklessness. In that case, there will be extenuating circumstances. Perhaps the sun was shining in her eyes. Hence, she is not to blame. Now, if an accident is blameless, it falls within the range of risks that are morally acceptable, irrespective of how harmful and tragic it is.

Importantly, this insight can be generalized without committing the anthropomorphic mistake. The idea is that an agent exhibits a control gap exactly if it poses risks to others that are morally unacceptable. And this level can differ between agents. In particular, the acceptable level of risk may be lower for artificial agents as compared to human beings. For example, we might forgive specific flaws that are part of the human condition, but we might not afford the same leniency to comparable deficiencies in artificial agents if they are not intrinsic to them. Thus, an agent exhibits a control gap precisely when it fails to meet the risk level that is deemed to be morally acceptable for that type of agent. It has not been suitably equipped to adequately respond to moral reasons or act as if they had been.

Thus, the control that an artificial agent possesses might be subpar. Perhaps an autonomous machine has not been properly designed or trained to handle certain unexpected events or circumstances. Suppose next that a collective agent has grown a lot in a short period of time. In the meantime, its management structure has become outdated. In such circumstances, artificial agents exhibit control gaps. As this reveals, control gaps in artificial agents are, in essence, due to faulty risk management (Hindriks and Veluwenkamp 2023; Hindriks 2024a). We go on to argue that this is to be avoided.[15]

5.2. Duty

The duty of moral control is the obligation to secure a proper level of moral control. Given the connection between control and risk, the argument for the existence of this duty is straightforward. Control gaps are deficiencies that lead to higher risk levels. As such, they are undesirable. Furthermore, they tend to give rise to blameful accidents, it is wrong for there to be control gaps. In light of this, we propose that there is a duty to ensure that there are no control gaps: to avoid or resolve them. Moral agents bear this duty themselves. Insofar as artificial agents are concerned, this duty is (also) borne by their enablers.

The moral control that an agent has depends on its constitution. What level is appropriate depends on context. Enablers such as designers, engineers and manufacturers are responsible for the constitution of an artificial agent. Regulators are responsible for whether and where they can be used. For instance, drones might have to keep a certain distance from built-up areas. Although regulators set the standards, manufacturers are often involved in testing as well. Furthermore, the conditions under which artificial agents operate are actively shaped by institutions, such as competition agencies (markets) and local and state governments (roads). Thus, the wide range of activities that fall under the duty of moral control concern the constitution of the artificial agent and the circumstances under which it acts.

Moral control is an ability. It is a prerequisite for moral action. As such, the duty of moral control is a meta-obligation. To see why, suppose the duty is not properly fulfilled. As a consequence, the artificial agent exhibits

[15]Mirzaeighazi and Stenseke (2024, 3) have criticized our account of control gaps because they take it to imply that, as soon as an artificial agent makes fewer mistakes than humans, it is admissible. But we are not committed to this. As just discussed, the standards that artificial agents have to meet may well be higher than those for human beings.

a shortfall in control. Because of this deficiency, it is prone to make mistakes and get into accidents. In relevant cases, they constitute violations of first-order obligations that the artificial agent has. And an agent should ensure that it is in a position to fulfill them. This reveals that an artificial agent that lacks proper moral control is prone to violate its first-order obligations in ways that are in principle preventable. This is why the duty of moral control is best seen as a second-order obligation, a duty to see to it that it is in a position to fulfill its first-order obligations.[16]

Thus far, we have argued that artificial agents can exhibit control gaps. Furthermore, they or their enablers have a duty to ensure that it has the proper level of moral control. We now go on to consider violations of this obligation.

5.3. Blame

A control gap is a deficiency in moral control. It increases the risk of wrongdoing. If an artificial agent exhibits a control gap, this will be due to a failure on the part of its enablers. They must have failed to fulfill their duty of moral control. In principle, they are (directly) to blame for this. Importantly, a control gap can exist without manifesting itself in an accident or mistake. So, an artificial agent can have a control gap without actually violating any of its obligations. In such cases, the control gap is latent. At the same time, an artificial agent with a control gap might malfunction because of this. If it does, the enablers are indirectly to blame for this.

As CAs do not only have enablers but also operators, the question arises whether and how control gaps affect their backward-looking responsibilities. Consider a corporate control gap that is due to bad management. This could lead, for instance, to a shortage in supplies or an overabundance of pressure. As a consequence, the operators might make mistakes. In some such way, a control gap forms an obstacle for operators to properly fulfill their tasks. Because that obstacle is beyond their control, they are thereby excused. Thus, the presence of a control gap explains why operators are blameless in such cases.

Strikingly, Copp (2007) and Pettit (2007) deny that, in the examples they consider, enablers bear responsibility for what the CA does. This is

[16]This proposal is inspired by the idea that an agent who has an obligation to do something should see to it that they are in a position to fulfill it (Goodin 2012). And it is somewhat analogous to the notion of 'a meta-task responsibility' that Van den Hoven (1998) introduced in the context of information technology, which was based on Goodin's earlier notion of a task-responsibility.

important to their argument, because it implies that none of the members is to blame for (contributing to) the corporate wrongdoing. As discussed above, they take this to imply that the collective agent must be blameworthy. But why are the operators blameless? Copp and Pettit provide explanations of this that differ between cases. What we add is a further explanation: if the responsibilities of the operators are defeated, chances are that this is due to a control gap. This is particularly likely if the enablers are to blame. In fact, if enablers are to blame while operators are blameless, this is a defeasible indicator of a corporate control gap.

6. How to close control gaps

Control gaps of artificial agents present a danger to society. So, they should be closed. By doing so, artificial agents can be accommodated in a society. But how can control gaps be closed? List claims that responsibility gaps can be resolved by securing full-blown moral agency for artificial agents. Insofar as CAs are concerned, the idea is as follows:

> Society, via its regulatory authorities, should permit the creation and operation of powerful group agents, such as corporations and other organizations in high-stakes settings, only if structures are in place to ensure their fitness to be held responsible for their corporate actions. (List 2021, 1229)

As this formulation reveals, this proposal is restricted to collective agents that are powerful. List also formulates the restriction in terms of 'high stake settings' that expose others to significant risks (List 2021, 1230).

List adds an important qualification when he formulates the constraint for AMs: they should be permitted to operate in high stake settings only if they, 'or at least their legal representatives,' can be held responsible for what they do (2021, 1230). Presumably, this qualification is meant to accommodate the reality of semi-autonomous machines. Now, the moral agency of the legal representatives of artifacts is hardly in question. So, if this is the solution, there is not really a problem. Furthermore, even if it resolves problems concerning the attribution of responsibility, it does not address what we regard as the core problem, that of a shortfall in control.

Securing full-blown moral agency is the obvious way to fulfill the duty of moral control (A). However, there is another one. Instead of increasing the control of the agent, this can also be done by decreasing the degree of control that is needed (B). Furthermore, each of these goals can be achieved in different ways. If the agent is a moral agent, the control

gap can be resolved by increasing the amount of moral control it has. This serves (A1) to secure full-blown moral agency. In contrast, if it is a minimal agent, the thing to do is (A2) to enable the artificial agent to better emulate moral control.

More modest solutions lower the requisite degree of control by decreasing the risks that the artificial agent poses. This can be done by (B1) re-engineering the agent or by (B2) adjusting the circumstances in which it operates. Thus, there are four ways of fulfilling the duty of moral control. We go on to discuss them in some detail. Here the comparison between AMs and CAs will turn out to be particularly fruitful.

(A1) Secure Moral Agency

A control gap can be closed by making sure that the artificial agent is a moral agent. To this end, minimal agents are to be transformed into moral agents. If the artificial agent is already a moral agent, its moral control is to be maintained at or restored to the appropriate level. The idea is that, by doing so, it acquires the requisite abilities to handle the degree of risk to which it exposes other agents. In this respect, they become comparable to human beings.[17]

(A2) Meaningful Human Control

As things are now, few if any machines possess full autonomy. Instead, they have a rather limited amount of moral control, if any at all. Furthermore, they still rely on operators to a significant extent. Together with its operators, a semi-autonomous agent can be seen as a system that can and should emulate moral control. To this end, it must be designed such that its operators possess 'meaningful human control.' If the operators can suitably interact with the semi-autonomous agent, they will be able to intervene and prevent accidents and other mistakes. Thus, the control gap of a semi-autonomous agent can be closed by enabling it to better emulate moral control.

In a range of situations, semi-autonomous agents decide and act on their own. To do so in a responsible way, they must track the moral reasons and values that apply in the context at issue. And it should be able to adjust its decisions and actions in their light. However, exactly because it is semi-autonomous, the way it does so will be deficient. In such cases, meaningful human control is meant to enable operators to improve the decisions or actions of the machines (Santoni de Sio and

[17] A moot question is what this means for the will of the artificial agent. Suppose it possesses full-blown moral agency. Does this imply that it can suffer from weakness of will? And what about a bad will? Should we be worried about killer robots after all?

Van den Hoven 2018). Such control goes beyond mere human supervision. The operators must understand, not only how it functions, but also how they can influence its functioning meaningfully, in the light of the relevant reasons and values. In this way, the decisions and actions of the semi-autonomous machine should be traceable back to a human operator or enabler, who is aware of this.

Although meaningful human control was proposed to close responsibility gaps, we adopt it as a way to close control gaps (Hindriks 2024a). Furthermore, we propose that its use is not restricted to machines but extends to CAs. To indicate how, it is useful to say more about what they are. A collective agent can be seen as a realized social structure (Ritchie 2020, Hindriks 2024b). Such a structure encompasses a system of roles (French 1984). And those roles are occupied by human beings, who are the members of the collective agent.

For such a structure to function as a moral agent, its members must be able to interact in such a way that the structure as a whole constitutes or emulates a moral agent. For this to be the case, their roles must be suitably aligned. If they are not, information that is passed on by one member to another might, for instance, never arrive where it needs to be. And projects might have to be abandoned because the requisite cooperation between those involved does not materialize. Thus, better aligning the roles can plausibly be seen as a way of enhancing meaningful human control within the collective agent.

Thus, control gaps can be closed by means of meaningful human control. Doing so provides for a second way to fulfill the duty of moral control.

(B1) Safety Engineering

The first two options, A1 and A2, increase the moral control that the artificial agent has or emulates. The remaining ones approach the problem from the opposite direction. They decrease the risk that the artificial agent poses to other agents. And they thereby reduce the moral control needed for it to function. This can be done by means of safety engineering (Sklet 2006; Van Nunen et al. 2018). In particular, the artificial agent or its operations can be surrounded by safety barriers.

Such barriers can be proactive or reactive. Proactive safety barriers are designed to prevent the occurrence of undesirable events. These can be thought of as the first line of defense against potential issues, working to preemptively avoid harmful situations. While proactive barriers aim to prevent incidents, reactive barriers serve to limit their impact. Hence,

they are implemented to control or mitigate the consequences of an event after it has occurred. The goal is to keep the harmful effects to a minimum if an undesired event were to occur.

Again, we believe that this proposal generalizes from AMs to CAs. A construction company might require those present on a construction site to wear safety helmets by equipping helmets with RFID (Radio-Frequency Identification) tags that interact with sensors at the entrance to a hazardous area. Companies might develop software with hard-coded limits to the financial risks employees can take. These measures are examples of proactive safety barriers. On the other hand, as a form of reactive safety barrier, backup generators or uninterruptible power supplies (UPS) can be installed to maintain operations and prevent data loss in the event of a power outage. Et cetera. Now, because these barriers decrease the risk posed to others, they also lower the requisite amount of moral control. Thus, it provides for a third way of satisfying the duty of moral control.

(B2) Social Control

The risks that an artificial agent poses to others can also be decreased by adjusting the conditions or circumstances in which it operates. As mentioned in section 5.2, drones might have to keep a certain distance from built-up areas. Similarly, an online casino, which heavily relies on algorithms, might have to be certified by a security testing company that ensures that your financial and personal information is safe and unhackable. In contrast to B1's focus on *technical* measures, the approach mentioned here requires *social* mechanisms, such as self-regulation within a social domain, or *legal* restrictions that are imposed from the outside.

A regulatory framework can encompass the entire lifecycle of AMs, from their design and deployment to their operation and eventual decommissioning. Furthermore, if they include safety measures, the approach is hybrid and combines elements of B1 and B2. In the extreme, social control can result in a legal straightjacket, which leaves little or no room for discretion on the part of those who deploy the AM or operate the CA. Thus, the duty of moral control can be fulfilled by means of social control.

7. Conclusion

Control gaps constitute a practical problem that requires urgent attention. In this respect, our approach differs from quietism, which fails to recognize that there is a practical problem. In contrast, fatalism overreacts

by assuming that risk levels are bound to be too high, such that the problem cannot be solved. We reject such a dystopian stance. Our realist alternative acknowledges that there are various ways in which control gaps can be closed and the level of risk can be made morally acceptable. Securing full-blown moral agency is just one of them.

Acknowledgments

We gratefully acknowledge the helpful feedback from Hein Duijf, Niels de Haan and Ann-Katrien Oimann.

Disclosure statement

The authors have no conflicts of interest to declare. All co-authors have seen and agree with the contents of the manuscript and there is no financial interest to report.

ORCID

Herman Veluwenkamp http://orcid.org/0000-0003-1783-459X
Frank Hindriks http://orcid.org/0000-0002-5818-4071

References

Bhargava, V. R., and M. Velasquez. 2019. "Is Corporate Responsibility Relevant to Artificial Intelligence Responsibility?" *Geo. JL & Pub. Pol'y* 17:829–851.

Braham, Matthew, and Martin van Hees. 2018. "Voids or Fragmentation: Moral Responsibility for Collective Outcomes." *The Economic Journal* 128 (612): F95–113. https://doi.org/10.1111/ecoj.12507.

Cavalcante Siebert, L., M. L. Lupetti, E. Aizenberg, N. Beckers, A. Zgonnikov, H. Veluwenkamp, D. Abbink, E. Giaccardi, G.-J. Houben, and C. M. Jonker. 2022. "Meaningful Human Control: Actionable Properties for AI System Development." *AI and Ethics* 3: 1–15.

Collins, Stephanie. 2019. "Collective Responsibility Gaps." *Journal of Business Ethics* 154:943–954. https://doi.org/10.1007/s10551-018-3890-6.

Copp, D. 2006. "On the Agency of Certain Collective Entities: An Argument from "Normative Autonomy"." *Midwest Studies in Philosophy* 30: 194–221.

Copp, D. 2007. "The Collective Moral Autonomy Thesis." *Journal of Social Philosophy* 38 (3): 369–388. https://doi.org/10.1111/j.1467-9833.2007.00386.x.

Duijf, H. 2022. *The Logic of Responsibility Voids*. Cham: Springer.

Fischer, J. M., and M. Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility (Issue 2)*. New York: Cambridge University Press.

French, P. A. 1984. *Collective and Corporate Responsibility*. Columbia University Press.

Goodin, R. E. 2012. "Excused by the Unwillingness of Others?" *Analysis* 72:18–24. https://doi.org/10.1093/analys/anr128.

Hakli, R., and P. Mäkelä. 2019. "Moral Responsibility of Robots and Hybrid Agents." *The Monist* 102 (2): 259–275. https://doi.org/10.1093/monist/onz009.

Hart, H. L. A. 2008. *Punishment and Responsibility: Essays in the Philosophy of Law*. Oxford: Oxford University Press.

Himmelreich, J. 2019. "Responsibility for Killer Robots." *Ethical Theory and Moral Practice* 22 (3): 731–747. https://doi.org/10.1007/s10677-019-10007-9.

Hindriks, F. 2024a. *The Moral Failures of Collective Agents: Responsibility Voids and Control Gaps*. Unpublished Manuscript.

Hindriks, F. 2024b. "The Social Construction of Collective Moral Agency." *Social Theory and Practice* 50 (3).

Hindriks, F., and H. Veluwenkamp. 2023. "The Risks of Autonomous Machines: From Responsibility Gaps to Control Gaps." *Synthese* 201 (1): 21. https://doi.org/10.1007/s11229-022-04001-5.

Horowitz, M., and P. Scharre. 2015. *Meaningful Human Control in Weapon Systems: A Primer*. Washington: Center for a New American Security.

Köhler, Sebastian, Neil Roughley, and Hanno Sauer. 2017. "Technologically Blurred Accountability?: Technology, Responsibility Gaps and the Robustness of Our Everyday Conceptual Scheme." In *Moral Agency and the Politics of Responsibility*, edited by Cornelia Ulbert, Peter Finkenbusch, Elena Sondermann, and Tobias Debiel, 51–68. London: Routledge.

Königs, Peter. 2022. "Artificial Intelligence and Responsibility Gaps: What Is the Problem?" *Ethics and Information Technology* 24 (3).

Laukyte, M. 2017. "Artificial Agents among Us: Should We Recognize Them as Agents Proper?" *Ethics and Information Technology* 19:1–17. https://doi.org/10.1007/s10676-016-9411-3.

List, C. 2021. "Group Agency and Artificial Intelligence." *Philosophy & Technology* 34 (4): 1213–1242. https://doi.org/10.1007/s13347-021-00454-7.

List, C., and P. Pettit. 2006. "Group Agency and Supervenience." *Southern Journal of Philosophy S1* 44: 85–105. https://doi.org/10.1111/j.2041-6962.2006.tb00032.x.

List, C., and P. Pettit. 2011. *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford University Press.

Matthias, A. 2004. "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata." *Ethics and Information Technology* 6 (3): 175–183. https://doi.org/10.1007/s10676-004-3422-1.

McKenna, M. 2008. "Putting the Lie on the Control Condition for Moral Responsibility." *Philosophical Studies* 139:29–37. https://doi.org/10.1007/s11098-007-9100-5.

Mirzaeighazi, S., and J. Stenseke. 2024. "Responsibility Before Freedom: Closing the Responsibility Gaps for Autonomous Machines." *AI and Ethics* 1 (1).

Munch, Lauritz, Jakob Mainz, and Jens Christian Bjerring. 2023. "The Value of Responsibility Gaps in Algorithmic Decision-Making." *Ethics and Information Technology* 25 (1): 21. https://doi.org/10.1007/s10676-023-09699-6.

Nyholm, Sven. 2023. "Responsibility Gaps, Value Alignment, and Meaningful Human Control Over Artificial Intelligence." In *Risk and Responsibility in Context*, edited by Adriana Placani and Stearns Broadhead, 191–213. Routledge.

Orwell, G. 2021. *Nineteen Eighty-Four*. Penguin Classics (Original work published 1949).
Pettit, P. 2007. "Responsibility Incorporated." *Ethics* 117 (2): 171–201. https://doi.org/10.1086/510695.
Pettit, P. 2017. "The Conversable, Responsible Corporation." *The Moral Responsibility of Firms* 1: 15–35. https://doi.org/10.1093/oso/9780198738534.003.0002.
Placani, Adriana, and Stearns Broadhead. 2024. *Risk and Responsibility in Context*. New York: Taylor & Francis.
Purves, D., R. Jenkins, and B. J. Strawser. 2015. "Autonomous Machines, Moral Judgment, and Acting for the Right Reasons." *Ethical Theory and Moral Practice* 18:851–872. https://doi.org/10.1007/s10677-015-9563-y.
Ritchie, K. 2020. "Social Structures and the Ontology of Social Groups." *Philosophy and Phenomenological Research* 100 (2): 402–424. https://doi.org/10.1111/phpr.12555.
Santoni de Sio, F., and J. Van den Hoven. 2018. "Meaningful Human Control Over Autonomous Systems: A Philosophical Account." *Frontiers in Robotics and AI* 15: 1–14.
Singer, A. E. 2013. "Corporate Moral Agency and Artificial Intelligence." *International Journal of Social and Organizational Dynamics in IT (IJSODIT)* 3 (1): 1–13. https://doi.org/10.4018/ijsodit.2013010101.
Sklet, S. 2006. "Safety Barriers: Definition, Classification, and Performance." *Journal of Loss Prevention in the Process Industries* 19 (5): 494–506. https://doi.org/10.1016/j.jlp.2005.12.004.
Sparrow, R. 2007. "Killer Robots." *Journal of Applied Philosophy* 24 (1): 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x.
Streumer, B. 2007. "Reasons and Impossibility." *Philosophical Studies* 136:351–384. https://doi.org/10.1007/s11098-005-4282-1.
Van den Hoven, M. 1998. "Moral Responsibility, Public Office and Information Technology." *Public Administration in an Information Age: A Handbook* 6: 97–112.
van de Poel, I. 2011. "The Relation Between Forward-Looking and Backward-Looking Responsibility." In *Moral Responsibility: Beyond Free Will and Determinism*, edited by N. A. Vincent, I. van de Poel, and J. van den Hoven, 37–52. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-1878-4_3.
Van Nunen, K., P. Swuste, G. Reniers, N. Paltrinieri, O. Aneziris, and K. Ponnet. 2018. "Improving Pallet Mover Safety in the Manufacturing Industry: A bow-tie Analysis of Accident Scenarios." *Materials* 11 (10): 1955. https://doi.org/10.3390/ma11101955.
Vargas, M. 2005. "The Trouble with Tracing." *Midwest Studies in Philosophy* 29:269–291. https://doi.org/10.1111/j.1475-4975.2005.00117.x.
Véliz, C. 2021. "Moral Zombies: Why Algorithms are not Moral Agents." *AI & Society* 36:487–497. https://doi.org/10.1007/s00146-021-01189-x.
Veluwenkamp, H. 2024. What Responsibility Gaps are and What They should Be. Unpublished Manuscript.
Wallace, R. J. 1994. *Responsibility and the Moral Sentiments*. Cambridge: Harvard University Press.
Yeung, K. 2019. *Responsibility and AI: Council of Europe Study DGI*. Council of Europe. https://rm.coe.int/responsability-and-ai-en/168097d9c5.